

A Modified Construction for a Support Vector Machine to Accommodate Class Imbalances

Matt Parker, Colin Parker

Abstract

Given a training set with binary classification, the Support Vector Machine identifies the hyperplane maximizing the margin between the two classes of training data. This general formulation is useful in that it can be applied without regard to variance differences between the classes. Ignoring these differences is not optimal, however, as the general SVM will give the class with lower variance an unjustifiably wide berth. This increases the chance of misclassification of the other class and results in an overall loss of predictive performance. An alternate construction is proposed in which the margins of the separating hyperplane are different for each class, each proportional to the standard deviation of its class along the direction perpendicular to the hyperplane. The construction agrees with the SVM in the case of equal class variances. This paper will then examine the impact to the dual representation of the modified constraint equations.

1 A Recap: The Classical SVM Construction

For Section 1, we follow the construction given by Hastie, Tibshirani, and Friedman in *The Elements of Statistical Learning* [3]. We will parallel this approach in Section 2 when constructing the alternate method.

Suppose we have training data consisting of pairs of observations and labels, (x_i, y_i) , for $i = 1, \dots, N$, with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. We may define a hyperplane by:

$$\{x : f(x) = x^T \beta + \beta_0 = 0\} \quad (1)$$

where β is a vector perpendicular to the hyperplane. An associated classification rule is induced by:

$$G(x) = \text{sign}[x^T \beta + \beta_0] \quad (2)$$

The goal of finding a separating hyperplane which maximizes the margin M for a linearly separable dataset, the minimum perpendicular distance to a datapoint of either class, can be formalized as:

$$\max_{\beta, \beta_0, \|\beta\|=1} M \quad (3)$$

$$\text{subject to } y_i(x_i^T \beta + \beta_0) \geq M \quad i = 1, \dots, N \quad (4)$$

This can be more conveniently rephrased by removing the requirement β be a unit vector, and setting $M = \frac{1}{\|\beta\|}$:

$$\min_{\beta, \beta_0} \|\beta\| \quad (5)$$

$$\text{subject to } y_i(x_i^T \beta + \beta_0) \geq 1 \quad i = 1, \dots, N \quad (6)$$

Now define slack variables $\zeta_i, i = 1, \dots, N$ by

$$\zeta_i = \max(0, 1 - y_i(x_i^T \beta + \beta_0)) \quad (7)$$

This gives us a framework to relax the assumption of linear separability. Noting that misclassifications occur when $\zeta_i > 1$, we see the slack variables are the proportion of the margin by which various points fall within their respective margins. We may control the amount of slack by imposing the additional condition:

$$\sum_{i=1}^N \zeta_i \leq \text{constant} \quad (8)$$

for some constant. This is computationally equivalent to the following expression:

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \zeta_i \quad (9)$$

$$\text{subject to } \zeta_i \geq 0, \quad y_i(x_i^T \beta + \beta_0) \geq 1 - \zeta_i \quad \forall i \quad (10)$$

where the parameter C replaces the constant in the previous expression. The corresponding Lagrange primal function is given by:

$$L_P = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \zeta_i - \sum_{i=1}^N \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \zeta_i)] - \sum_{i=1}^N \mu_i \zeta_i \quad (11)$$

which is to be minimized with respect to β, β_0 , and ζ_i . Setting the respective derivatives equal to zero, we get the equations:

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i \quad (12)$$

$$0 = \sum_{i=1}^N \alpha_i y_i \quad (13)$$

$$\alpha_i = C - \mu_i \quad \forall i \quad (14)$$

and positivity constraints $\alpha_i, \mu_i, \zeta_i \geq 0 \forall i$. By substituting the above three equations into the Lagrangian dual we obtain the Wolfe dual, given by:

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'} \quad (15)$$

$$= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} \langle x_i, x_{i'} \rangle \quad (16)$$

In addition, the Karush-Kuhn-Tucker conditions yield:

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \zeta_i)] = 0 \quad (17)$$

$$\mu_i \zeta_i = 0 \quad (18)$$

$$y_i (x_i^T \beta + \beta_0) - (1 - \zeta_i) \geq 0 \quad (19)$$

for $i = 1, \dots, N$. These equations collectively uniquely define the solution to the dual problem.

2 A Modified Approach: Accommodating Difference in Class Variance

The original construction of the SVM for linearly separable data has the goal of maximizing the margin $M = \frac{1}{\|\beta\|}$. In the event of a noticeable difference between class variances in the direction of β (perpendicular to our separating hyperplane), the SVM ends up positioning the decision boundary closer to the class with larger variance [say, class A] than would be optimal. The new construction accommodates these class imbalances by increasing the margin of the class of greater variance.

It will be useful at this point to define a few terms. For class K , element $x_j \in K$, and separating hyperplane $\{x : x^T \beta + \beta_0 = 0\}$, define $\sigma_{K,\beta} = \sigma_{y_j, \beta}$ to be the standard deviation of elements of class K in the direction of β :

$$\sigma_{K,\beta} = \sigma_{y_j,\beta} = \text{Var}(\{(x_i - \bar{x}) \cdot \left(\frac{\beta}{\|\beta\|}\right) \mid i \in K\})^{\frac{1}{2}} \quad (20)$$

$$= \left(\sum_{j:y_j=y_i} \left[(x_j - \bar{x}) \cdot \left(\frac{\beta}{\|\beta\|}\right) \right]^2 \right)^{\frac{1}{2}} \quad (21)$$

and, for class K and arbitrary hyperplane $\{x : x^T \beta + \beta_0 = 0\}$, define the margin of class K to be:

$$M_K = \min_{x_i \in K} y_i \left(\frac{x_i^T \beta + \beta_0}{\sigma_{y_i,\beta}} \right) \quad (22)$$

We will now seek to find the separating hyperplane which maximizes $\min_K M_K$, the minimum margin over all classes. As an aside, a byproduct of the classic construction of the SVM yields the equality $M_A = M_B$ when separating classes A and B , since the maximum margin is obtained when the separating hyperplane is midway between both classes. Our new construction will yield as a byproduct the equality:

$$\frac{M_A}{\sigma_{A,\beta}} = \frac{M_B}{\sigma_{B,\beta}} \quad (23)$$

This shows that in the event our classes have equal variance in the direction of β , the modified construction coincides with the classical SVM.

3 Examining Implications to Dual Representation

Maximizing $\min_K M_K$ modifies the optimization problem to the pair of equations:

$$\min_{\beta, \beta_0} \|\beta\| \quad (24)$$

$$\text{subject to } y_i \left(\frac{x_i^T \beta + \beta_0}{\sigma_{y_i,\beta}} \right) \geq 1 \quad i = 1, \dots, N \quad (25)$$

Slightly redefining slack variables according to the fraction of the respective margins they span yields:

$$\zeta_i = \max \left(0, 1 - y_i \left(\frac{x_i^T \beta + \beta_0}{\sigma_{y_i,\beta}} \right) \right) \quad (26)$$

and the corresponding modified SVM equations are given by:

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \zeta_i \quad (27)$$

$$\text{subject to } \zeta_i \geq 0, \quad y_i \left(\frac{x_i^T \beta + \beta_0}{\sigma_{y_i, \beta}} \right) \geq 1 - \zeta_i \quad \forall i \quad (28)$$

We can now formulate the corresponding Lagrangian (primal) function as:

$$L_P = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \zeta_i - \sum_{i=1}^N \alpha_i \left[y_i \sigma_{y_i, \beta}^{-1} (x_i^T \beta + \beta_0) - (1 - \zeta_i) \right] - \sum_{i=1}^N \mu_i \zeta_i \quad (29)$$

which we again minimize with respect to β, β_0 , and ζ_i . Setting derivatives with respect to β_0 and ζ_i equal to zero, we get similar results:

$$0 = \sum_{i=1}^N \alpha_i y_i \sigma_{y_i, \beta}^{-1} \quad (30)$$

$$\alpha_i = C - \mu_i \quad \forall i \quad (31)$$

and a slightly more complex equation when doing the same with respect to β :

$$0 = \nabla_{\beta} L_P \quad (32)$$

$$= \nabla_{\beta} \left(\frac{1}{2} \|\beta\|^2 - \sum_{i=1}^N (\alpha_i y_i) \left(\sigma_{y_i, \beta}^{-1} \right) (x_i^T \beta + \beta_0) \right) \quad (33)$$

$$= \beta - \sum_{i=1}^N \alpha_i y_i x_i \sigma_{y_i, \beta}^{-1} + \sum_{i=1}^N (\alpha_i y_i) \left(\sigma_{y_i, \beta}^{-2} \right) (x_i^T \beta + \beta_0) (\nabla_{\beta} \sigma_{y_i, \beta}) \quad (34)$$

Expanding $\sigma_{y_i, \beta}$ to its representation in (21), we may utilize the Hadamard product notation \circ and the fact

$$\nabla_{\beta} \left((x_j - \bar{x}) \cdot \left(\frac{\beta}{\|\beta\|} \right) \right) = (x_j - \bar{x}) \cdot \left(\frac{\|\beta\|^2 - \beta \circ \beta}{\|\beta\|^3} \right) \quad (35)$$

where \circ is the Hadamard product, to obtain:

$$\begin{aligned}
0 = & \beta - \sum_{i=1}^N \alpha_i y_i x_i \sigma_{y_i, \beta}^{-1} + \\
& + \sum_{i=1}^N \left[\alpha_i y_i \sigma_{y_i, \beta}^{-3} (x_i^T \beta + \beta_0) \left(\sum_{j: y_j = y_i} \left[(x_j - \bar{x}) \cdot \left(\frac{\beta}{\|\beta\|} \right) \right] \left[(x_j - \bar{x}) \left(\frac{\vec{\mathbf{1}} \|\beta\|^2 - \beta \circ \beta}{\|\beta\|^3} \right) \right] \right) \right]
\end{aligned} \tag{36}$$

where $\vec{\mathbf{1}}$ is the vector of ones $[1, \dots, 1]$.

This gives us a working representation of the equivalent dual optimization equations under the new construction, and a forthcoming paper will be examining the solvability of the above in general in light of the other constraint equations, as well as consequent impacts to kernelizability of the method. We will also examine in depth the circumstances in which our alternate construction outperforms a traditional Support Vector Classifier, and attempt to quantify them.

References

- [1] Trevor Hastie, Robert Tibshirani, and Jerome Freidman. *The Elements of Statistical Learning*. Springer-Verlag, New York, New York, 2009.
- [2] Andrew Ng. *CS229 Lecture Notes*. [<http://cs229.stanford.edu/notes/cs229-notes3.pdf>]
- [3] Robert Gunn, *Support Vector Machines for Classification and Regression*. Technical Report for University of Southampton, Southampton, England, 1998.